# Seeing Through .onion: Multimodal Information Extraction and Category Discovery in Tor Markets

Zhouguo Chen[†§], Chunmian Wang[*†‡], Leran Ren[†], Ming Yang[†]
[†] School of Computer Science and Engineering, Southeast University, Nanjing, China
czgexcel@163.com, chunmianwang@fyust.edu.cn,{leranren, yangming2002}@seu.edu.cn
[‡] School of Computing and Artificial Intelligence, Fuyao University of Science and Technology, Fuzhou, China
[§] The 30th Research Institute of China Electronics Technology Group Corporation, Chengdu, China

***Abstract*—The anonymity of the Tor dark web has fostered numerous hidden online marketplaces. These markets typically present products using multimodal information, including images and text. Sellers often pair real images of illegal goods with simplified, vague, or inaccurate text descriptions. However, existing dark web market analyses generally rely on single-modal text classification or isolated Natural Language Processing (NLP) tasks, and thus fall short of delivering a comprehensive understanding. In this paper, we propose a multimodal framework, DARKINTELLIGENCE, for holistic market analysis. First, DARKINTELLIGENCE leverages a LoRA-fine-tuned vision-language model to process the multimodal product information, generating unified and comprehensive records of both product details and seller profiles. Next, it applies Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to adaptively cluster products and construct a hierarchical taxonomy. A large language model then assigns human-interpretable labels to each cluster. The rich information extracted by DARKINTELLIGENCE enables accurate and explainable analysis of dark web markets from multiple perspectives, including product structure, inventory, pricing, and seller profiles. We evaluate DARKINTELLIGENCE on a dataset of 27,726 multimodal entries collected from five realistic and representative Tor markets. The framework achieves 94.30% accuracy in product field information extraction and effectively discovers more hidden categories.**

***Index Terms*—Tor Markets, Large Language Model, Information Extraction, Category Discovery**

## I. INTRODUCTION

Tor [1] is currently the most widely used anonymous communication system, providing end-to-end anonymity for both service providers and clients through multi-hop routing and onion encryption mechanism. These anonymous services (i.e., hidden services) which operate within the Tor network ensure that both users and service providers remain anonymous, creating a highly private and secure dark web environment. The dark web constitutes a complex and dynamic ecosystem, with online markets playing a pivotal role in its structure. These markets are not only centers of illicit activities but also offer a wide array of goods and services that are typically difficult to obtain or trace on the clearnet.

Understanding dark web markets and their product offerings is crucial, as it helps uncover patterns in supply, market trends, and seller behaviors. By analyzing the ecological distribution of these markets, researchers can gain deeper insights into the operational dynamics of these platforms, identify emerging threats, and formulate more targeted intervention strategies. The main challenge lies in effectively mapping this ecosystem and extracting actionable intelligence from the vast and concealed data within these markets.

Existing research in dark web intelligence, particularly in content classification and information extraction, remains fragmented and insufficient for a comprehensive understanding of marketplace ecosystems. Early approaches rely on single-modal text models, while newer multimodal techniques (such as CLIP [2], unsupervised multimodal clustering [3], and context-based image similarity [4]) are mostly developed on open-domain data and do not address challenges specific to dark-web content, including noisy screenshots, domain-specific jargon, and cross-lingual product pages. These methods also assume predefined label sets, limiting their ability to discover fine-grained categories in a data-driven manner. For intelligence extraction, traditional NLP pipelines [5], [6], [7], [8] and LLM-based methods [9] focus on sentence- or document-level semantics and do not generate structured, marketplace-level records. Although recent advances in multimodal extraction [10], [11], [12], [13] and cross-lingual embeddings [14] enhance representation learning, current work still lacks unified frameworks that integrate multimodal representation, adaptive category discovery, and query-oriented intelligence extraction for large-scale dark-web analysis.

To address these limitations, we adopt a unified, market-centric approach that jointly tackles structured information extraction, adaptive categorization, and cluster-based intelligence analysis for dark-web markets. On the extraction side, we move beyond purely text-based pipelines and treat marketplace pages as multimodal objects. Screenshots are first processed with object detection to localize concrete products (e.g., drug packages, counterfeit documents, weapons). Optical Character Recognition (OCR)[15] is applied to text-rich regions. A Low-Rank Adaptation (LoRa) fine-tuned vision–language model converts these visual and textual cues into normalized product- and seller-level records. For classification, instead of relying on a fixed, manually designed label set, we perform adaptive category discovery in the embedding space and then use a large language model to assign concise, human-

* Corresponding author: Chunmian Wang of Fuyao University of Science and Technology, China.

interpretable labels to the discovered clusters, yielding an explainable taxonomy that aligns with marketplace semantics. Building on these structured representations and interpretable categories, we further organize products into clusters and conduct group-level statistical analysis, enabling high-level market intelligence (such as seller-product relationships, total value of products across different categories) to be generated directly from the underlying multimodal data.

In this paper, we propose DARKINTELLIGENCE, a multi-modal framework for systematically mapping and analyzing dark-web markets. Our contributions are as follows.

- we propose a structured information extraction pipeline that combines object detection, OCR, and a LoRA-fine-tuned multimodal large language model to convert noisy screenshots and page text into unified product- and seller-level records, covering fields such as names, descriptions, prices, and activity indicators.
- we introduce a multimodal category discovery approach that embeds product descriptions with BGE-M3 and applies hierarchical Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to uncover a three-level taxonomy of market products, enabling adaptive, fine-grained categorization across heterogeneous sites.
- We perform a cluster-based intelligence analysis that profiles each discovered group in terms of inventory, pricing, and seller behavior. We evaluate it on 27,726 cleaned entries collected from five Tor markets and demonstrate high accuracy in extraction and categorization while remaining efficient enough for practical dark-web intelligence analysis.

## II. BACKGROUND

We first introduce the Tor network and the dark web in §II-A, and then we present Large Language Models (LLMs) in §II-B, which enable efficient extraction and analysis of textual and visual data vital for understanding dark-web content.

### A. Tor and Dark Web

Tor is an overlay network that ensures user anonymity by routing traffic through volunteer-operated nodes. As shown in Figure 1, when a user accesses a clearnet service, the Tor client establishes a three-hop circuit, encrypting the traffic at each stage to prevent any node from knowing both the origin and destination. Tor also supports hidden services with ".onion" domains, which allow users and service providers to remain anonymous, enhancing privacy.

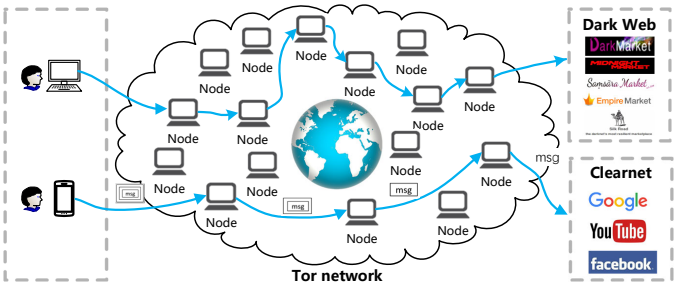
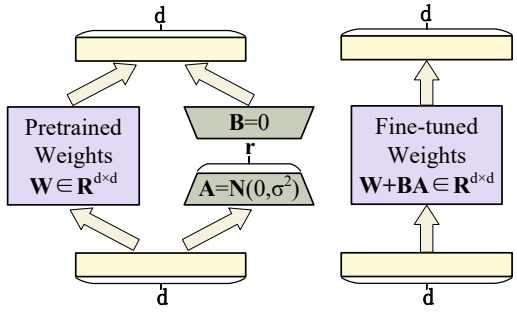



Fig. 1: Overview of the Tor Network




(a) LoRA fine-tuning (b) LoRA inference

Fig. 2: LoRA optimization

The high level of anonymity provided by Tor makes the dark web an ideal environment for both legitimate applications, such as protecting freedom of speech, and illicit activities, including illegal trade, data selling, and cybercrime. A prominent feature of the Tor dark web is the emergence of online marketplaces, which mirror the structure of conventional e-commerce platforms but typically facilitate the trade of illegal goods and services. These markets often include categories such as narcotics, firearms, counterfeit documents, stolen data, and hacking tools. They adopt familiar mechanisms such as product listings, vendor profiles, customer ratings, and cryptocurrency-based payment systems, usually relying on Bitcoin or Monero.

### B. Large Language Model and LoRA Fine-Tuning

LLMs are designed to predict the next token in a sequence based on the context of previous tokens. These models are typically autoregressive, generating output one token at a time, with each new token being conditioned on both the input prompt and previously generated tokens. This mechanism allows LLMs to capture complex language patterns and dependencies, making them highly effective for tasks such as text generation, classification, and information extraction.

LoRA is an efficient fine-tuning method for adapting pre-trained large language models to specific tasks. Instead of retraining the entire model, LoRA updates only a small set of parameters—two low-rank matrices, $\mathbf{A}$ and $\mathbf{B}$. As shown in Figure 2, during the fine-tuning phase (Figure 2(a)), the pre-trained weight matrix $\mathbf{W}$ is combined with a low-rank matrix $\mathbf{A}$ (initialized with random values) while keeping matrix $\mathbf{B}$ set to zero. The parameter $r$, which defines the rank of the update, is much smaller than the original dimension $d$, resulting in a more computationally efficient fine-tuning process. Only the matrices $\mathbf{A}$ and $\mathbf{B}$ are updated, reducing memory and computational overhead. At inference time (Figure 2(b)), the fine-tuned weights are computed by adding the product of $\mathbf{A}$ and $\mathbf{B}$ to the pre-trained weight matrix $W$. This merged weight matrix $\mathbf{W} + \mathbf{BA}$ is then used for inference, eliminating any additional overhead and ensuring efficient performance without introducing extra latency.

## III. DARKINTELLIGENCE DESIGN AND IMPLEMENTATION

We first present an overview of the DARKINTELLIGENCE framework in §III-A, then describe structured information

extraction and multimodal category discovery in §III-B and §III-C, respectively. Finally, we show intelligence analysis in §III-D.

### A. Overview

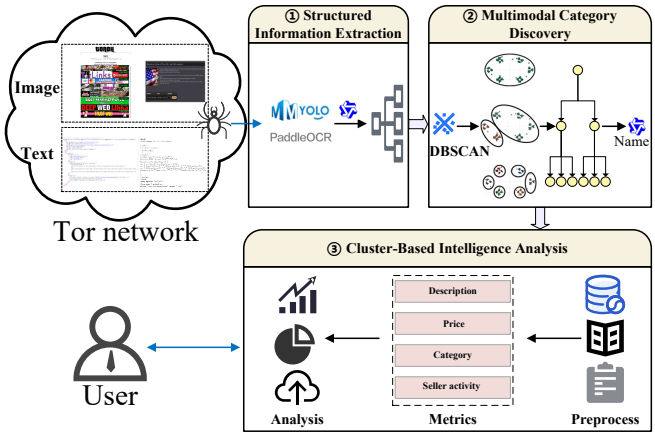



Fig. 3: The DARKINTELLIGENCE framework

The DARKINTELLIGENCE framework presents a comprehensive approach to dark-web product analysis (see Figure 3), integrating structured information extraction, multimodal category discovery, and cluster-based intelligence extraction. Initially, the framework employs object detection, OCR recognition, and fine-tuned multimodal large language models to extract product metadata and seller information from dark-web pages. By utilizing the Qwen3-VL-2B model for image analysis and integrating tools such as YOLOv8 [16] for object detection and PaddleOCR for text-heavy image extraction, the framework ensures high precision in capturing product attributes, including names, descriptions, prices, and seller details. Next, the framework introduces multimodal category discovery using hierarchical DBSCAN clustering. This method organizes products into a multi-layer structure, categorizing them into bottom, middle, and top layers based on granularity. By applying the BGE-M3 model for sentence embeddings, DBSCAN clusters product descriptions to uncover hidden market structures, with large language models generating meaningful labels for each category. Finally, the framework analyzes each cluster's features, such as product distribution and seller activity, to identify market patterns. These integrated methods provide a robust and scalable solution for dark-web intelligence gathering.

### B. Structured Information Extraction

We propose a method for extracting dark-web market product metadata and seller information using object detection, OCR recognition, and fine-tuning a multimodal large language model. *Product metadata* includes attributes such as product name, description, category, and price, while *seller information* encompasses details like seller nickname, launch time, product reviews, and other relevant data.

We first perform object detection using state-of-the-art techniques (such as YOLOv8) to identify and localize relevant objects within dark-web screenshots. These objects might include products (e.g., narcotics or weapons). Object detection facilitates the segmentation of these images into defined regions of interest, which are then processed further for detailed analysis. Next, OCR is applied to the text-heavy regions of the images, particularly for items such as counterfeit documents or text-based product descriptions. PaddleOCR is employed to extract textual data with high precision, which is essential for accurately capturing product names, descriptions, and prices.

After extracting visual and textual information, a fine-tuned multimodal LLM (i.e., Qwen2.5-VL-7B) is employed to integrate both modalities. The model, which is trained using LoRA, is specifically adapted to handle the noisy, often incomplete data typical of dark-web marketplaces. In this process, only the low-rank matrices of the model are fine-tuned, minimizing computational overhead while retaining the model's overall capabilities. This adaptation enables the LLM to process and structure the extracted data into normalized product-level records, encompassing fields such as product names, descriptions, prices, categories, and seller information.

### C. Multimodal Category Discovery

We propose a hierarchical DBSCAN clustering approach to identify categories in heterogeneous market product data, providing an adaptive and robust framework that sheds light on the latent structure of dark-web product categories. In this method, we divide the product categories into three levels: *bottom layer*, *middle layer*, and *top layer*. The hierarchical clustering process progressively merges categories from fine-grained subcategories to broader, coarse-level categories, improving classification accuracy and interpretability.

Let $S = \{s_1, s_2, \ldots, s_n\}$ represent the set of $n$ sentences (or product descriptions) extracted from the dark web marketplace. Each sentence $s_i$ is mapped into a 1024-dimensional vector space using the BGE-M3 model, ensuring alignment of heterogeneous features from structured text data. The vector representation of the $i$-th sentence, $v_i$, is given by Equation (1).

$$v_i = BGE(s_i) \tag{1}$$

The similarity between two sentences $s_i$ and $s_j$ is computed using the cosine similarity measure:

$$sim(s_i, s_j) = \frac{v_i \cdot v_j}{\|v_i\| \|v_j\|}, \tag{2}$$

where $v_i \cdot v_j$ is the dot product of the sentence vectors and $\|v_i\|$ and $\|v_j\|$ are the Euclidean norms of the respective vectors. This measure quantifies how similar two sentences are in terms of their contextual meaning in the unified vector space.

Then, We use the DBSCAN algorithm to perform hierarchical clustering across three levels: bottom, middle, and top layers. DBSCAN requires two key parameters: $\epsilon$ and $minPts$. $\epsilon$ is the radius parameter that defines the maximum distance between two points for them to be considered part of the same neighborhood, and $minPts$ is the minimum number of points required to form a dense region (i.e., a cluster). For each level of clustering, we set different values of $\epsilon$ and $minPts$ to control the granularity of the clusters. The bottom-up aggregation

strategy ensures consistency in category assignments. Finally, we use a LLM (i.e., Qwen3-VL-2B) to generate meaningful, concise labels for the final clusters by summarizing the content of each major category. This step leads to the creation of a hierarchical classification structure for dark web products.

### D. Cluster-Based Intelligence Analysis

In this section, we introduce a cluster-based intelligence analysis approach that utilizes structured data from product listings and clusters identified through DBSCAN. This analysis focuses on profiling product attributes and seller behaviors to gain insights into dark web market dynamics.

**Seller Behavior Profiling.** We calculate the *number of product releases* for each seller by aggregating product listings. For each seller $S_i$, the number of releases $N(S_i)$ is computed as:

$$N(S_i) = \sum_{j=1}^{n} \delta(S_j, S_i), \tag{3}$$

where $\delta(S_j, S_i)$ is 1 if product $j$ belongs to seller $S_i$, and 0 otherwise. This quantifies seller activity. Next, we compute the *total monetary value* of each category by summing the prices of products in that category. The total value $V(C_m)$ for category $C_m$ is:

$$V(C_m) = \sum_{k \in C_m} P_k, \tag{4}$$

where $P_k$ is the price of product $k$.

**Market Concentration Analysis.** To assess *market concentration*, we use the Cumulative Distribution Function (CDF) to analyze the relationship between the number of sellers and product releases. This helps identify whether a small number of sellers dominate the market, indicating potential monopolistic behavior.

## IV. Evaluation

In this section, we first introduce the environment setup and dataset in §IV-A. Then we present the experimental methods and results for structured information extraction, multimodal classification discovery, and cluster-based intelligence analysis in §IV-B, §IV-C and §IV-D, respectively.

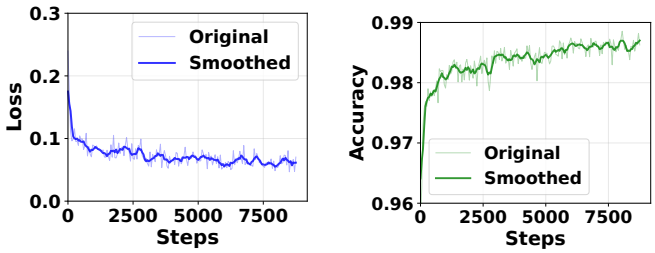
### A. Environmental Setup

We conduct the experiments on a high-performance computing server equipped with four NVIDIA RTX A6000 GPUs and an Intel Xeon-class CPU with 1 TB of system memory.

**Dataset.** The raw data collects from five dark web markets comprises 33,501 product screenshots. After data cleaning and deduplication, 27,726 valid entries are retained. Table I summarizes the per-site statistics before and after cleaning.

TABLE I: Dataset Construction: Raw vs. Cleaned Screenshots

| Site Name | Collection Period | Raw(#) | Cleaned(#) |
|---|---|---|---|
| White House | 25/04/12-25/04/25 | 22,015 | 18,078 |
| Chang'an Nightless City | 25/04/12-25/04/14 | 10,203 | 8,500 |
| Zion | 25/04/14-25/04/14 | 686 | 574 |
| BMG | 25/04/12-25/04/12 | 388 | 365 |
| LolitaCity | 25/04/14-25/04/14 | 209 | 209 |




(a) Training Loss (b) Token Accuracy

Fig. 4: Training Loss and Accuracy During Model Training

### B. Structed Information Extraction

**Methodology.** To evaluate the performance of structured information extraction, we randomly selecte and manually annotate 7,296 samples from the cleaned dataset and fine-tune the model Qwen3-VL-2b on a benchmark dataset. We visualize loss and accuracy to evaluate convergence and performance improvement during training. The loss value reflects the magnitude of error in the training process, and the accuracy measures the proportion of correctly predicted samples among all samples.

In addition, we evaluate the model's performance in extracting information from text using *Field Extraction Accuracy* (denoted as $F$). Specifically, we test the model's ability to extract the following fields: product name, category, price, seller username, and release time. *Field Extraction Accuracy* measures the proportion of correctly extracted fields across all samples in the test set. Given a test set with $M$ samples, where $m$ represents the number of samples with correctly extracted fields, *Field Extraction Accuracy* $F$ is derived as:

$$F = \frac{m}{M}. \tag{5}$$

**Result.** Figure 4(a) and Figure 4(b) display the model loss and accuracy during 8,757 steps of training, respectively. The loss curve fluctuates significantly at the onset. Then it decreases consistently as training proceeds, indicating gradual learning and optimization. The accuracy curve increases continuously, rising from 96% to 98.8%, which indicates that the model consistently captures data patterns.

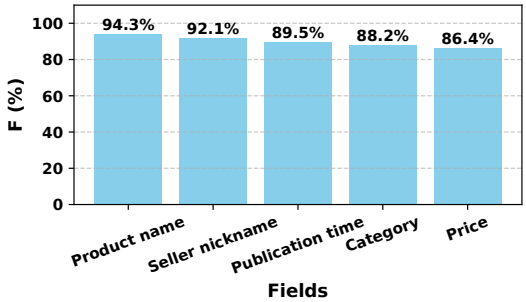



Fig. 5: Field Extraction Accuracy

Figure 5 shows high accuracy in extracting key product fields, with the best performance for *product name* (94.3%) and *seller nickname* (92.1%). Accuracy for *publication time* and *category* is lower (89.5% and 88.2%, respectively), indicating some extraction challenges. *Price* has the lowest

accuracy (86.4%), likely due to variations in formatting and non-standardized data.

### C. Multimodal Category Discovery

**Methodology.** To perform multi-level classification of dark web products, we begin by selecting representative features from the extracted product information. Specifically, we use the product name and description as the primary representations of each product, which are then concatenated into a single sentence. These concatenated sentences serve as the input for generating product embeddings. We utilize the BGE-M3 model to generate vector representations for each sentence, effectively capturing the semantic meaning of the product descriptions. Once these embeddings are obtained, we apply the DBSCAN algorithm to cluster the product sentences into distinct categories. In the DBSCAN clustering process, we set the $minPts$ parameter to 2, which ensures that each group consists of at least two products. The clustering granularity is controlled by adjusting the neighborhood radius parameter $\epsilon$. Products that form clusters with only a single item are treated as discrete groups, indicating that these products are relatively independent.

**Result.** Figure 6 demonstrates that DBSCAN clustering results in progressively broader product classifications as the radius $\epsilon$ increases. At a radius of 0.2, the clustering algorithm identifies 617 clusters with 2,855 discrete products. This radius captures a finer level of detail, with a relatively large number of smaller, more distinct product categories, which we designate as the *bottom layer*. Increasing the radius to 0.3 results in 48 clusters, with 373 products classified as discrete, indicating a moderate level of granularity where more products are grouped into fewer categories. This level of classification corresponds to the *middle layer*, where the categories become more generalized, and products share broader semantic similarities. At the highest radius of 0.4, only 7 clusters are formed, comprising 34 discrete products, suggesting a coarser classification level. This larger radius merges several previously distinct categories, indicating broader, more overarching product groups, which we classify as the *top layer*. The number of clusters for different layers is shown in Table II.

TABLE II: Product Categories at Different Layers

| Layers | $\epsilon$ | Clusters (#) | Discrete Points (#) |
|---|---|---|---|
| Top Layer | 0.4 | 7 | 34 |
| Middle Layer | 0.3 | 48 | 373 |
| Bottom Layer | 0.2 | 617 | 2855 |

### D. Intelligence Analysis

**Methodology.** We conduct a detailed analysis of the extracted product information using structured extraction and adaptive category discovery methods. For this purpose, we focus on the popular Chang'an Nightless City market as a case study. A total of 8,500 products are used for information extraction and classification. This dataset is analyzed to explore product attributes, seller behaviors, and market trends, allowing for a comprehensive examination of dark web market dynamics.

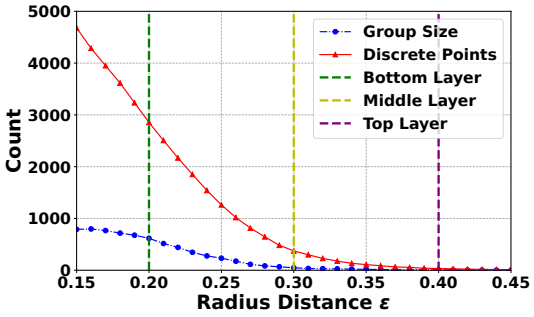



Fig. 6: DBSCAN Clustering Results

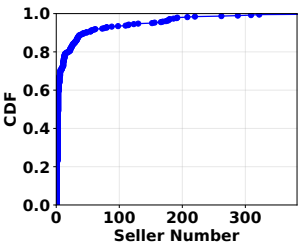



Fig. 7: Seller and Product Quantity Distribution

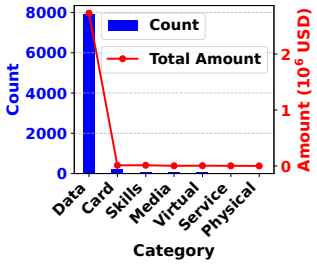



Fig. 8: Product Categories and Total Value Distribution

**Results.** Among the 8,500 products, we identify a total of 382 distinct sellers and classify the products into 7 main categories. Figure 7 shows the CDF of sellers and product quantities. The results reveal that just 27% of the sellers (the top 105 sellers by product count) account for 80% of the total product listings. This indicates a high concentration of product listings among a small group of dominant sellers, suggesting potential market monopolization within certain product categories.

Furthermore, we perform a classification of the products into their respective categories. Figure 8 illustrates the distribution of products across the 7 main categories, along with their total monetary values. Notably, category *Data* dominate the marketplace, comprise 93.56% of the total products, with a total value of $2,730,010. Other categories account for much smaller shares both in terms of product quantity and total value.

## V. Discussion

**Limitation.** This study is limited to five major dark web marketplaces, primarily focusing on English and Chinese content. Expanding to other platforms and languages is essential for broader applicability. The framework's performance is influenced by the quality of the collected data, which may contain noise or inconsistencies. Additionally, clustering results depend on the vector models (i.e., BGE-M3) and product descriptions, so future work should focus on fine-tuning these models and improving description accuracy for better clustering.

**Ethical Considerations.** Due to the high anonymity of the Tor network, collecting data from dark web marketplaces does not compromise the privacy of the sites. Additionally, the data collection process is done at a slow rate, ensuring that it does not impose any additional load or pressure on the service sites. We adhere to data minimization principles, collecting only the necessary data related to illicit activities while ensuring the protection of user privacy.

## VI. Related Work

**Content Classification.** Early content classification methods focus primarily on single-modal text-based approaches. Radford *et al.* [2] introduce CLIP, which maps images and text into a shared space via contrastive learning, enabling natural language-based image classification and zero-shot learning. Zhang *et al.* [3] propose unsupervised multimodal clustering for semantic mining. For dark web content classification, Prado-Sanchez *et al.*[17] compare eight large language models under zero-shot conditions for illegal content, identifying DeepSeek Chat, Grok, and Gemini 2.0 Flash as top performers. Additionally, Vignoli and Monteiro [18] conduct a comparative study of deep and dark web content to aid in classification.

**Intelligence Extraction.** The extraction of intelligence from dark web content initially relies on single-modal NLP techniques. Lample *et al.* [5] develop NER methods, while Cabot and Navigli [6] work on relation extraction, and Lu *et al.* [7] focus on event extraction, solve the problems of traditional decomposition strategies in an end-to-end manner, build upon Cowie and Lehnert's [8] foundational work. Wang *et al.* [9] demonstrate the potential of large language models (LLMs) for NER tasks. With the advancement of multimodal techniques, Guo *et al.* [10] provide a survey of multimodal extraction approaches, and Yang *et al.* [11] introduce the DEEPEVAL benchmark test to explore the ability of LMMs in understanding the deep semantics of images. Shukla *et al.* [12] propose a large multimodal model named PATENTLMM for generating descriptions of patent drawings, while Luo *et al.* [13] develop bilingual fine-tuning methods. Chen et al. [14] propose a novel text embedding model called M3 Embedding, which demonstrates excellent generality in multiple languages, functionalities, and granularities, supporting semantic retrieval in over 100 languages.

## VII. conclusion

This paper introduces a multimodal DarkIntelligence framework for dark web marketplace intelligence that jointly addresses structured information extraction, adaptive category discovery, and cluster-based analysis. By combining visual-centric extraction, hierarchical DBSCAN clustering, and LLM-assisted label generation, our approach maps noisy, heterogeneous Tor market pages into interpretable product taxonomies and structured records. Experiments on a cleaned dataset of 27,726 entries from five Tor markets show that the framework achieves high accuracy in field extraction and product classification, while discovering meaningful latent categories and supporting low-latency analysis.

## Acknowledgment

This work was supported by the National Key Research and Development Program of China under Grants 2023YFB3106600 and 2023YFC3605800, the Jiangsu Provincial Key R&D Program under Grant BE2022065-5, the Jiangsu Provincial Key Laboratory of Network and Information Security under Grant BM2003201, the Key Laboratory of Computer Network and Information Integration of the Ministry of Education of China under Grant 93K-9, and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

## References

[1] Tor Project, "Tor." https://www.torproject.org/, 2025.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

[3] H. Zhang, H. Xu, F. Long, X. Wang, and K. Gao, "Unsupervised multimodal clustering for semantics discovery in multimodal utterances," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 18–35, 2024.

[4] V. Franzoni, A. Milani, S. Pallottelli, C. H. C. Leung, and Y. Li, "Context-based image semantic similarity," in *Proceedings of the 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pp. 1280–1284, 2015.

[5] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, pp. 260–270, 2016.

[6] P. L. H. Cabot and R. Navigli, "Rebel: Relation extraction by end-to-end language generation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2370–2381, 2021.

[7] Y. Lu, H. Lin, J. Xu, X. Han, J. Tang, A. Li, L. Sun, M. Liao, and S. Chen, "Text2event: Controllable sequence-to-structure generation for end-to-end event extraction," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pp. 2795–2806, 2021.

[8] S. Sarawagi *et al.*, "Information extraction," *Foundations and Trends® in Databases*, vol. 1, no. 3, pp. 261–377, 2008.

[9] S. Wang, X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, G. Wang, and C. Guo, "Gpt-ner: Named entity recognition via large language models," in *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4257–4275, 2025.

[10] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63373–63394, 2019.

[11] Y. Yang, Z. Li, Q. Dong, H. Xia, and Z. Sui, "Can large multimodal models uncover deep semantics behind images?," in *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1898–1912, 2024.

[12] S. Shukla, N. Sharma, M. Gupta, and A. Mishra, "Patentlmm: Large multimodal model for generating descriptions for patent figures," in *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI)*, pp. 20488–20496, 2025.

[13] L. Luo, J. Ning, Y. Zhao, Z. Wang, Z. Ding, P. Chen, W. Fu, Q. Han, G. Xu, Y. Qiu, D. Pan, J. Li, H. Li, W. Feng, S. Tu, Y. Liu, Z. Yang, J. Wang, Y. Sun, and H. Lin, "Taiyi: a bilingual fine-tuned large language model for diverse biomedical tasks," *Journal of the American Medical Informatics Association*, vol. 31, no. 9, pp. 1865–1874, 2024.

[14] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," in *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, 2024.

[15] PaddlePaddle Team, "Paddleocr 3.0." https://www.paddleocr.ai/main/index.html, 2025.

[16] R. Varghese and M. Sambath, "Yolov8: A novel object detection algorithm with enhanced performance and robustness," in *Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pp. 1–6, 2024.

[17] V.-P. Prado-Sánchez, A. Domínguez-Díaz, L. De-Marcos, and J.-J. Martínez-Herráiz, "Zero-shot classification of illicit dark web content with commercial llms: A comparative study on accuracy, human consistency, and inter-model agreement," *Electronics*, vol. 14, no. 20, 2025.

[18] R. G. Vignoli and S. D. Monteiro, "Deep web and dark web: Similarities and dissiparities in the context of information science," *Journal of Information Science*, 2020.